Efficient LLM Comparative Assessment: A Product of Experts Framework for Pairwise Comparisons

Anonymous ACL submission

Abstract

LLM-as-a-judge approaches are a practical and effective way of assessing a range of text tasks. However, when using pairwise comparisons to rank a set of candidates, the computational cost scales quadratically with the number of candidates, which has practical limitations. This paper introduces a Product of Expert (PoE) framework for efficient LLM Comparative Assessment. Here individual comparisons are considered experts that provide information on a pair's score difference. The PoE framework combines the information from these experts to yield an expression that can be maximized with respect to the underlying set of candidates, and is highly flexible where any form of expert can be assumed. When Gaussian experts are used one can derive simple closed-form solutions for the optimal candidate ranking, as well as expressions for selecting which comparisons should be made to maximize the probability of this ranking. Our approach enables efficient comparative assessment, where by using only a small subset of the possible comparisons, one can generate score predictions that correlate well with human judgements. We evaluate the approach on multiple NLG tasks and demonstrate that our framework can yield considerable computational savings when performing pairwise comparative assessment. With many candidate texts, using as few as 2% of comparisons the PoE solution can achieve similar performance to when all comparisons are used.

1 Introduction

The advent of instruction-following (Wei et al., 2021; Ouyang et al., 2022) Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023) has enabled systems to exhibit impressive zero-shot capabilities on a range of Natural Language Processing (NLP) tasks. One such practical application is in Natural Language Generation (NLG) evaluation (Fabbri et al., 2021), where LLMs can be prompted to assess the quality of texts for particular attributes (Wang et al., 2023; Liu et al., 2023a; Zheng et al., 2023). A popular approach is LLM comparative assessment, where pairwise comparisons are used to determine which of two texts is better (Zheng et al., 2023; Qin et al., 2023; Liusie et al., 2024b). Although using pairwise comparisons has shown to better align with human preferences (Liusie et al., 2024b) than LLM scoring approaches (Wang et al., 2023; Liu et al., 2023a), the set of all comparisons scales quadratically with the number of inputs, which can be impractical in real-world use cases. Therefore, one may instead consider methods that only use a subset of comparisons to predict the scores, such that performance is maintained in computationally efficient settings.

Due to its applicability to sports, search and many other domains, the task of going from a subset of comparisons to a final ranking/scoring has been well-studied and extensively explored (Davidson and Farquhar, 1976; David, 1963; Luce, 2005; Cattelan, 2012). However, in the majority of setups, the comparative decisions are binary (win/loss, although occasionally also win/loss/tie). LLMs, however, not only provide the outcome of the comparison but also additional information, such as the associated probability that A is better than B. Despite this available information, current LLM comparative works often leverage naive metrics such as win-ratio (Qin et al., 2023; Zheng et al., 2023; Liusie et al., 2024b) and average probability (Park et al., 2024; Molenda et al., 2024), with little analysis on how to maximally extract the information from the comparisons.

This paper introduces a theoretical framework for viewing comparative assessment that enables practical scoring even in cases when the full set of comparisons is not used. We conceptualize the process as a Product of Experts (PoE) (Hinton, 1999; Welling, 2007), where each comparative decision is assumed to provide information on the quality difference between the two competing texts. The framework is highly flexible and can use any form of expert. By considering two forms of experts, namely 1) the Gaussian distribution with linear assumptions and 2) an extension of the Bradley-Terry (BT) model for soft probabilities (motivated by looking at its limiting behaviour), we demonstrate that the PoE framework for comparative assessment can achieve efficient and effective NLG assessment. With the Gaussian expert, the framework yields a closed-form solution for the scores, which conveniently yields standard metrics when using the full set of comparisons. We demonstrate that our Product of Expert framework leads to significant performance boosts across models, datasets and assessment attributes, and even when using a fraction of the possible comparisons, can achieve high performance with minimal performance degradation from the full set.

This paper makes several contributions. 1) We introduce the PoE perspective of comparative assessment, a highly flexible theoretical framework which enables one to directly model the distribution of scores given a set of comparisons. 2) We propose two experts, a soft Bradley-Terry expert (by considering the limiting behaviour of BT) and a Gaussian expert that has closed-form solutions and can be used to select the most informative comparisons. 3) We demonstrate practically that the PoE solution yields significant computational savings and empirically show that convergence is reached significantly faster than when using other baseline approaches for several datasets.

2 Background and Related Work

Traditional/Tailored NLG Evaluation: Initially, the outputs of NLG systems were evaluated against ground-truth human-annotated references, using N-gram overlap metrics (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) or similarity metrics (Zhang et al., 2019). For more fine-grained evaluation, later studies developed bespoke evaluators for particular task dimensions such as summary consistency (Wang et al., 2020; Manakul et al., 2023; Kryściński et al., 2020) or dialogue coherence (Dziri et al., 2019; Ye et al., 2021). Further extensions considered unified evaluators, which evaluate multiple independent attributes (Mehri and Eskenazi, 2020; Yuan et al., 2021; Zhong et al., 2022). A drawback with these traditional NLG evaluation approaches is that they typically are bespoke towards particular tasks and attributes and, therefore, cannot easily be extended to new domains.

LLM-Based NLG Evaluation: Given the impressive instruction-following (Ouyang et al., 2022; Chung et al., 2022) capabilities of LLMs such as GPT-4 (Achiam et al., 2023) and open-sourced variants (Chung et al., 2022; Touvron et al., 2023), recent works have studied leveraging these LLMs for general zero-shot NLG evaluation. Methods include GPTScore (Fu et al., 2023), which computes the LLM likelihood of generating the response, and LLM-as-a-judge approaches (Zheng et al., 2023) that prompt models to provide scores (Wang et al., 2023; Kocmi and Federmann, 2023; Liu et al., 2023a) or use pairwise comparisons to determine which of two responses is better (Qin et al., 2023; Liusie et al., 2024b).

LLM Comparative Assessment: Various recent works have used pairwise LLM comparative assessment for ranking texts: Liusie et al. (2024b) demonstrate that for moderate-sized LLMs, comparative assessment outperforms LLM scoring as well as various bespoke baselines. They compute the win-ratio using all $N(N-1)$ comparisons as well as with a subset of comparisons (where large degradations are observed). Further, Qin et al. (2023) use pairwise comparisons for retrieving relevant sources, both using the full set of comparisons as well as sorting-based algorithms. Park et al. (2024) apply comparative assessment to dialogue evaluation, computing the average probability over a randomly sampled set of comparisons as the score quality. They also adapt the model with supervised training. Lastly, Liu et al. (2024) demonstrate limitations for LLM scoring and, therefore, instead, consider pairwise comparisons. They introduce PAirwise-preference Search (PAIRS), a variant of the merge sort algorithm using LLM probabilities.

Comparisons to Scores: Although LLMs have only recently been used as pairwise evaluators, the problem of ranking a set of candidates from a set of pairwise comparisons has been extensively studied in many different contexts, including sports (Beaudoin and Swartz, 2018; Csató, 2013), information retrieval (Cao et al., 2007; Liu et al., 2009) and social studies (Manski, 1977; Louviere et al., 2000). Arguably the most widely used parametric model is the Bradley-Terry model (Bradley and Terry, 1952), which models the win probabilities based on the difference of the latent scores of the compared items. The latent scores are deduced by maximizing the likelihood of the observed pairwise comparison data, with various works discussing algorithms that converge to the solution (Davidson and Farquhar, 1976; David, 1963; Cattelan, 2012). Additionally, (Chen et al., 2022) investigate predicting rankings under the Bradley-Terry-Luce model (Luce, 2005), while TrueSkill (Herbrich et al., 2006; Minka et al., 2018) extends the Bradley-Terry model to incorporate uncertainties in player skills (in a sports context) under a Bayesian framework.

3 A Product of Experts Perspective of Comparative Assessment

Let $x_{1:N} \in \mathcal{X}$ be a set of N candidate texts and $s_{1:N} \in \mathbb{R}$ the scores of the texts for a particular assessed attribute. Given a set of K pairwise comparisons, $\mathcal{C}_{1:K}$, the objective is to determine a predicted set of scores, $\hat{s}_{1:N}$, that are close to the true scores, $s^*_{1:N}$.

3.1 The Bradley–Terry Model

For traditional comparative assessment set-ups, outcomes are usually discrete and either binary (win/loss) or ternary (win/draw/loss). A standard approach of going from a set of discrete comparisons $\mathcal{C}_{1:K}$ to predicted scores $\hat{s}_{1:N}$ is the Bradley–Terry model (Bradley and Terry, 1952; Zermelo, 1929). Assuming each comparison C_k is of the form (i, j, y_{ij}), where $y_{ij} \in \{0, 1\}$ represents whether x_i is better than x_j, one can adopt a probabilistic binomial model where the probability of victory depends solely on the difference of scores, $\mathtt{P}(y_{ij}|s_i–s_j) = \sigma(s_i–s_j)$. The most popular form is the sigmoid function, $\sigma(x) = 1/(1 + e^{-x})$. The Bradley-Terry model treats the scores as parameters of the model, and aims to maximize the likelihood of the observations,

$$\mathtt{P}(\mathcal{C}_{1:K}|s_{1:N}) = \prod_{i,j\in\mathcal{C}_{1:K}} \mathtt{P}(y_{ij}|s_{1:N}) \qquad (1)$$

$$\mathtt{P}(y_{ij}|s_{1:N}) = \sigma(s_i - s_j)^{y_{ij}} (1 - \sigma(s_i - s_j))^{1-y_{ij}} \qquad (2)$$

$$\hat{s}_{1:N} = \arg\max_{s_{1:N}} \mathtt{P}(\mathcal{C}_{1:K}|s_{1:N}) \qquad (3)$$

Although no closed-form solution exists, Zermello's algorithm (Zermelo, 1929) can be used to iterate the solution until convergence is reached. Furthermore, while Zermello's algorithm is known to be slow to converge (Dykstra, 1956; Hunter, 2004), later improvements have demonstrated faster convergence rates (Newman, 2023).

3.2 A Product of Experts Perspective

For LLM comparative assessment, as opposed to traditional binary comparative decisions, one has access to richer information, including the associated probability of a decision. Each comparison outcome can therefore be extended to the form (i, j, p_{ij}) where $p_{ij} = \mathtt{P}_{\mathtt{lm}}(y_i > y_j|x_i, x_j)$, the LLM probability of the comparative decision. To conveniently incorporate the soft-probability observations, we explore directly modelling the probability of scores given the comparative observations and reformulate the scores as a Product of Experts. A Product of Experts (PoE) (Hinton, 1999; Welling, 2007) combines the information gained from many individual experts by taking their product and normalizing the result. One can consider each comparison as information gained from independent experts, enabling the probability for the scores to be written as:

$$\mathtt{p}(s_{1:N}|\mathcal{C}_{1:K}) = \frac{1}{Z} \prod_{i,j\in\mathcal{C}_{1:K}} \mathtt{p}(s_i - s_j|C_k) \qquad (4)$$

Each expert can be conditioned on the observed LLM probability such that $\mathtt{p}(s_i - s_j|C_k) = \mathtt{p}(s_i - s_j|p_{ij})$. As a possible expert, we consider a form related to the limiting behaviour of the Bradley-Terry Model and re-express Equation 2 with a probabilistic classification result form,

$$\mathtt{p}(s_i - s_j|p_{ij}) = \frac{1}{Z_{ij}} \sigma(s_i - s_j)^{p_{ij}} (1 - \sigma(s_i - s_j))^{1-p_{ij}}$$

Where $0 < p_{ij} < 1$, and $Z_{ij} = \pi/\sin(p_{ij}\pi)$ is a normalization constant to ensure a valid probability density function. However, the experts are not restricted to sigmoid-based modelling; one can select any family of probability distributions, such as Gaussian experts, which are discussed next.

3.3 Properties of Gaussian Experts

Having Gaussian experts yields convenient properties in the PoE framework, such as a closed-form expression for the solution (Zen et al., 2011). If the underlying distribution is assumed to be Gaussian with the mean $f_\mu(p_{ij})$ and variance $f_\sigma(p_{ij})$ only dependent on the comparative probability, such that $\mathtt{p}(s_i - s_j|p_{ij}) = \mathcal{N}\big(s_i - s_j; f_\mu(p_{ij}), f_\sigma(p_{ij})\big)$, then by representing the scores in vector form, $\mathbf{s} = [s_{1:N}]$, one can express the distribution as,

$$\mathtt{p}(\mathbf{Ws}|\mathcal{C}_{1:K}) = \mathcal{N}\left(\mathbf{Ws}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right) \qquad (5)$$

Where $\mathbf{W} \in R^{K\times N}$ (illustrated in Appendix A.1) is a matrix representing the set of comparisons, such that for the k^{th} comparison between i and j $\mathbf{W}_{ki} = 1$, $\mathbf{W}_{kj} = -1$, and $\mathbf{W}_{km} = 0 \ \forall m \neq i, j$, $\mathbf{s}$ is the N-dimensional column vector of $s_{1:N}$, $\boldsymbol{\mu} \in R^K$ is a vector of the means, and $\boldsymbol{\sigma^2} \in R^K$ equivalently represents the variances,

$$\boldsymbol{\mu} = [f_\mu(p_{ij}^{(1)}), f_\mu(p_{ij}^{(2)}), ... f_\mu(p_{ij}^{(K)})]^\mathsf{T} \quad (6)$$

$$\boldsymbol{\sigma}^2 = [f_\sigma(p_{ij}^{(1)}), f_\sigma(p_{ij}^{(2)}), ... f_\sigma(p_{ij}^{(K)})]^\mathsf{T} \quad (7)$$

Note that as defined, the matrix $\mathbf{W}$ is not full rank since any shift of the scores $\boldsymbol{s}$ will yield an equivalent output. To address this, an additional expert on the first element can be added, such that $\mathtt{p}(s_1|\mathcal{C}_0) = \mathcal{N}(0, \sigma_0^2)$, prepending an extra row to all of $\mathbf{W}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$, yielding $\tilde{\mathbf{W}}$, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}$ respectively. The distribution takes a similar form, $\mathtt{p}(\tilde{\mathbf{W}}\mathbf{s}|\mathcal{C}_{1:K}) = \mathcal{N}(\tilde{\mathbf{W}}\mathbf{s}; \tilde{\boldsymbol{\mu}}, \text{diag}(\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}))$, which can be rearranged to yield a Gaussian expression for the score distribution, $\mathtt{p}(s_{1:N}|C_{1:K}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_s^*, \tilde{\boldsymbol{\Sigma}}_s^*)$, with mean and covariance matrix defined as,

$$\boldsymbol{\mu}_s^* = \tilde{\mathbf{W}}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} \quad (8)$$

$$\tilde{\boldsymbol{\Sigma}}_s^* = (\tilde{\mathbf{W}}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})^{-1} \quad (9)$$

where $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\tilde{\boldsymbol{\sigma}}^{\mathbf{2}})$ (the rearranging is shown in Appendix A.5). Therefore, the mean of the Gaussian provides a simple and closed-form solution to the maximum probability solution, $\hat{s}_{1:N}$,

$$\hat{\mathbf{s}} = \underset{s_{1:N}}{\arg\max}\, \mathtt{p}(s_{1:N}|\mathcal{C}_{1:K}) \quad (10)$$

$$= (\tilde{\mathbf{W}}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} \quad (11)$$

3.4 Further Gaussian Assumptions

A drawback with the Gaussian Expert is that producing $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}$ requires knowledge of both $f_\mu(p)$ and $f_\sigma(p)$. This is not available without human-annotated data, making the approach impractical for zero-shot applications. To enable a practical solution applicable in zero-shot settings, one can make two assumptions on the Gaussian experts: 1) that the variance is constant regardless of the predicted probability $f_\sigma(p) = \sigma^2$, and 2) that the mean scales linearly with the probability $f_\mu(p) = \alpha \cdot (p - \beta)$. These assumptions appear reasonable for several models and datasets (in Appendix Figure 10) and simplify the solution to,

$$\hat{\mathbf{s}} = \alpha \cdot (\tilde{\mathbf{W}}^\mathsf{T}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^\mathsf{T}\tilde{\boldsymbol{\mu}} \quad (12)$$

where $\tilde{\boldsymbol{\mu}}^\mathsf{T} = [0, p_{ij}^{(1)} - \beta, ..., p_{ij}^{(K)} - \beta]$. Note that a sensible choice might be $\beta = 0.5$, since when inputting texts of equal quality into an unbiased system, an average output probability of 0.5 would be expected. Further, the value of α only influences the relative spacing and subjective scale used to score the texts and can arbitrarily be set to 1.

3.5 Modelling Bias in Non-Symmetric Settings

LLMs can have inconsistent outputs where $p_{ij} \neq (1 - p_{ji})$ and, in particular, demonstrate positional bias (Zheng et al., 2023; Chen et al., 2024; Liusie et al., 2024a). Positional bias occurs when the system prefers one position over another such that $\mathbb{E}_{\mathtt{p}_{\text{lm}}(p)}[p] \neq 0.5$, while for unbiased systems, the expectation should be near 0.5. Combining the probabilities from both permutations such that $\tilde{p}_{ij} = \frac{1}{2} \cdot (p_{ij} + (1 - p_{ji}))$ ensures that $\tilde{p}_{ij} = (1 - \tilde{p}_{ij})$ and eliminates positional bias; however, it requires two LLM calls per comparison and may not be the best use of LLM calls. To efficiently minimize the impact of positional bias without requiring both LLM permutation calls, we investigate directly modelling model position bias into the experts. A simple approach is to introduce a bias parameter γ that shifts the experts such that, $P_\gamma(s_i - s_j|p_{ij}) = P(s_i - s_j - \gamma|p_{ij})$. The value of γ can be determined by noting that the expected score difference between two randomly sampled texts is zero, $\mathbb{E}[s_i - s_j] = 0$. For the linear Gaussian expert, this is equivalent to applying a linear shift in the mean, and therefore by considering $\mathcal{N}\big(s_i - s_j; \alpha \cdot (p_{ij} - \beta), \sigma^2\big)$,

$$\mathbb{E}[s_i - s_j] = \mathbb{E}[f_\mu(p_{ij})] = \alpha\big(\mathbb{E}[p_{ij}] - \beta\big) \quad (13)$$

setting the expression to zero yields that the debiasing term $\beta = \mathbb{E}[p_{ij}]$. For Bradley-Terry, though it can be shown that $f_\mu(p_{ij}) = -\pi \cdot \cot(\pi p_{ij})$, this value tends to infinty when p approaches either 0 or 1. Therefore, instead of setting the expected value of the skill difference for any random pair to be zero, we approximate finding the bias by ensuring the mode of the underlying (log-) distribution is 0 when the skill difference is 0. Based on this approximation, the resulting bias parameters for the extended Bradley-Terry is $\gamma = -\mathtt{logit}(\mathbb{E}[p_{ij}])$ (see Appendix A.8 for further details).

3.6 Comparison Selection

The previous theory detailed how to determine the predicted scores $\hat{s}_{1:N}$ given a random set of observed comparisons $\mathcal{C}_{1:K}$. As an extension, one may consider how to select the set of comparisons that provide the most information. Under the Gaussian model, the probability of the most likely set of scores is given as,

$$\mathtt{p}(\hat{s}_{1:N}|\mathcal{C}_{1:K}) = \frac{\sqrt{\mathtt{det}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}})}}{(2\pi\sigma^2)^{N/2}} \quad (14)$$

shown in Appendix A.5. For a fixed number of comparisons K, one may therefore aim to find the matrix $\tilde{\mathbf{W}}^*$ that minimizes the uncertainty,

$$\tilde{\mathbf{W}}^* = \arg\max_{\tilde{\mathbf{W}}} \mathtt{p}(\hat{s}_{1:N}|\mathcal{C}_{1:K}) \quad (15)$$

$$\equiv \arg\max_{\tilde{\mathbf{W}}} \mathtt{det}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}) \quad (16)$$

This can be approximated through an iterative greedy search. Assume that $\tilde{\mathbf{W}}^{(k)*}$ is the selected comparison matrix using k comparisons and $\mathbf{A}^{(k)*} = (\tilde{\mathbf{W}}^{(k)*\mathsf{T}}\tilde{\mathbf{W}}^{(k)*})^{-1}$. The next selected comparison $(\hat{\imath}, \hat{\jmath})$ can be calculated as,

$$\hat{\imath}, \hat{\jmath} = \arg\max_{i,j} \mathbf{A}^{(k)*}_{ii} + \mathbf{A}^{(k)*}_{jj} - 2 \cdot \mathbf{A}^{(k)*}_{ij} \quad (17)$$

As shown in Appendix A.6, where it is also shown that the inverse matrix $\mathbf{A}^{(k+1)*}$ can be updated efficiently from $\mathbf{A}^{(k)*}$.

4 Experimental Setup

4.1 Datasets

We consider a range of NLG evaluation datasets which have available ground-truth scores. For summary evaluation we use **SummEval** (Fabbri et al., 2021) which has 100 articles each with 16 machine-generated summaries evaluated on coherency (`COH`), consistency (`CON`), fluency (`FLU`), and relevancy (`REL`). For dialogue response generation we use **TopicalChat** (Mehri and Eskenazi, 2020) which has 60 dialogue contexts with six responses per context assessed on coherency (`COH`), continuity (`CNT`), engagingness (`ENG`), and naturalness (`NAT`). For question difficulty ranking, we use **CMCQRD** (Mullooly et al., 2023), which has 658 multiple-choice reading comprehension questions annotated on question difficulty. Lastly, for story evaluation, we use **HANNA** (Chhun et al., 2022) which has 1056 machine-generated stories annotated by humans on coherency (`COH`), complexity (`CMP`) and surprisingness (`SUR`). For CMCQRD and HANNA we compare the texts across all 658/1056 texts.

4.2 Methodology

Base Large Language Models Three different families of opensourced LLMs are used as judge LLMs: FlanT5 (3B, 11B) (Chung et al., 2022), instruction-tuned Mistral (7B) (Jiang et al., 2023) and Llama2-chat (7B, 13B) (Touvron et al., 2023).

LLM Pairwise Probability Calculation To get comparative probabilities, we follow Liusie et al. (2024b) and use $\mathtt{P(A)/(P(A)+P(B))}$. The symmetric set-up (where both permutations are done) is used unless stated otherwise, though in Section 5.4 the non-symmetric set-up is investigated.

Comparison Selection When considering comparative assessment with a subset of comparisons, the base experiments use a randomly drawn set of comparisons such that each comparison is equally likely to be chosen. For a set of inputs $x_{1:N}$, we randomly select K unique pairs (x_i, x_j) to be judged by the LLM, ensuring that each text x_i is involved in at least one comparison. Experiments begin with $K = 2N$ comparisons and K is incremented to the full set of comparisons, $K = N \cdot (N-1)$.

Scoring Methods Several different methods of mapping a set of comparisons to scores are used in this paper, categorized into binary decision-based or probability-based. For binary decision methods, our first baseline is the **win-ratio** which calculates the number of comparisons won as the quality score, as used in Qin et al. (2023); Liusie et al. (2024b); Raina and Gales (2024). The second baseline is the Bradley-Terry model, **BT**, (Bradley and Terry, 1952), where the solution is found by Zermelo (Zermelo, 1929) with a convergence threshold of $1e^{-4}$. Since any candidate that wins/loses all games will have an infinite score, a prior of $1/(N-1)$ wins is added to each selected comparison. For the methods that leverage the LLM probabilities, the baseline is the average probability **avg-prob** of a text in all its comparisons, as used in Park et al. (2024); Molenda et al. (2024). To better leverage the probabilistic information, our paper proposes to decompose the probability into a product of experts. We propose two variants; 1) **PoE-BT** which uses a variant of the Bradley-Terry model extended to soft probabilities (described in Section 3.2), and 2) **PoE-g** which uses the Gaussian expert with the linear mean and constant variance assumptions (described in Section 3.4). Lastly, the final method is **PoE-g-hard**, which applies the POE-gaussian framework, however, using hard binary decisions and not the soft probabilities.

Evaluation For SummEval and TopicalChat, the summary-level Spearman score is used as the assessment metric. For each context, we do pairwise comparisons using the LLM on the full set of $N(N-1)$ comparisons. We then simulate using a subset of comparisons by randomly selecting K of these outcomes. This process is repeated 100 times for a particular number of total comparisons, K, and we calculate both the mean and standard deviation of performance over the entire dataset. For Hanna and CMCQRD, there is no context dependence and therefore the number of candidate texts is much larger, with $N=1050$ and $N=550$ respectively. As such as we sample 200,000 comparisons (all symmetric), which is only a subset of the total possible comparisons, and provide analysis by simulating randomly sampling further subsets of these comparisons. For each K, we run 20 independent runs and average performance. For both datasets, equivalent tables for Pearson are provided in Appendix C.

5 Results

5.1 SummEval and TopicalChat

In this Section, we investigate whether the Product of Experts framework can yield performance boosts for SummEval and TopicalChat in efficient settings. SummEval has 16 candidates per context ($N=16$) and therefore considering all possible comparisons takes 240 comparisons, which though feasible, can be quite costly. Table 1 presents SummEval performance when only a subset of the comparisons are made, with the average Spearman rank correlation coefficient (SCC) over all contexts and attributes presented for different base LLMs. Equivalent tables for TopicalChat are provided in Appendix C.2 where similar trends are seen. The following observations can be made:

Average probability performs better than the win-ratio in efficient settings When considering the full set of comparisons ($K=240$) the performance of average probability is only marginally better than using win-ratio (within 1 SCC). However, when using 20% of the comparisons ($K=48$) the average probability yields significant gains of 3-4 SCC. This highlights that especially when only using a subset of comparisons, leveraging the soft probabilistic information is beneficial.

The PoE solution yields large gains in efficient settings Even when only using hard decisions, for $K=48$, both the Bradley-Terry model (BT) and the PoE Gaussian with hard decisions (PoE-g-hard) have mild performance gains over the win-ratio.

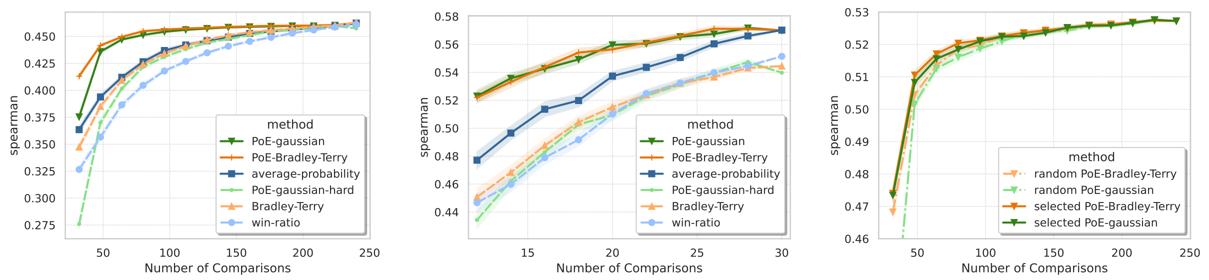
		Decisions		Probabilities		
System	K	Win-r	BT	Avg-pr	PoE-BT	PoE-g
Llama2-7B	48	21.6	23.4	24.0	26.8	26.6
	240	27.8	27.9	28.4	28.4	28.4
Llama2-13B	48	30.8	33.1	33.7	37.7	37.3
	240	39.3	39.3	39.3	39.3	39.3
Mistral-7B	48	29.7	31.9	31.1	33.2	32.8
	240	38.1	38.1	37.7	37.7	37.7
FlanT5-3B	48	34.1	36.6	38.4	42.6	42.4
	240	43.6	43.6	44.3	44.3	44.3
FlanT5-11B	48	31.2	33.4	34.7	38.5	38.4
	240	40.0	40.0	40.5	40.5	40.5

Table 1: Spearman Correlations for SummEval, averaged over all attributes (COH, CON, FLU, REL). K is the number of comparisons made, where $K=240$ is the full set of comparisons.

Nevertheless, the real benefits are seen when using PoEs with soft probabilities, with both POE-BT and PoE-g significantly outperforming the average probability. With these methods, when using only 20% of the comparisons, one can achieve performance close to when using the full comparison set (in four out of five cases within 2 SCC), when winratio would have degredations of up to 10 SCC. The findings are general and hold across the different SummEval attributes and models.

Gaussian PoE and BT PoE result in similar performing solutions When using full-comparisons, the Gaussian PoE solution can be shown to be equivalent to the average probability (shown in Appendix A.3) however the BT PoE approach will lead to a different solution. Nonetheless, the performance for both PoE-BT and PoE-g are very comparable for most models/datasets, in both the hard and soft set-ups. Further the Gaussian solution has the benefit of having a convenient closed form solution.

Convergence rates The results in Table 1 showed performance for the arbitrary chosen operating point of $K=48$. Figures 1a and 1b show the performance for two models/attributes while sweeping K from $K=N$ to the full set of comparisons, $K=N(N-1)/2$. The curves show that the performance improves smoothly while increasing number of comparisons, with the convergence rates considerably better with the PoE methods. Further plots for other models/tasks are provided in Appendix C.3.



(a) Llama2-13B, SummEval `REL`. (b) FlanT5-11B, TopicalChat, `ENG`. (c) FlanT5-3B, SummEval `COH`, **selected vs random**

Figure 1: Efficiency curves when sweeping K, the number of comparisons per context, where at each K the comparisons are randomly drawn 100 times. Average performance with 95% confidence is displayed.

5.2 Comparison Selection

The previous results used random comparisons, however, an alternative would be to pre-select a set of comparisons that maximizes the information gained from a fixed number of comparisons. Section 3.6 discusses how for the Gaussian-POE, this can be achieved with a practical greedy approximation. Table 2 illustrates that at the operating point of $K = 48$, pre-selecting the comparisons can provide further performance boosts, with the average performance of the probabilistic PoE approaches consistently increasing by 0.5 SCC for all approaches, at no extra cost. Although the theory was derived using the Gaussian assumptions, the performance boosts are seen for all methods, with the largest gains for the win-ratio. Lastly, Figure 1c shows that performance gains are significant when few comparisons are made, but as the number of comparisons grows, the performance difference between random and optimal selection is negligible.

System	Method	Win-r	Avg-pr	PoE-BT	PoE-g
Llama2-7B	Random	21.6	24.0	26.8	26.6
	Selected	23.0	24.5	27.3	27.2
Llama2-13B	Random	30.8	33.7	37.7	37.3
	Selected	32.4	34.6	38.2	38.0
Mistral-7B	Random	29.7	31.1	33.2	32.8
	Selected	31.4	32.2	34.0	33.9
FlanT5-3B	Random	34.1	38.4	42.7	42.4
	Selected	36.0	39.3	43.2	42.9
FlanT5-11B	Random	31.2	34.7	38.4	38.4
	Selected	33.1	35.7	39.2	39.0

Table 2: SummEval Spearman correlations when using the greedy optimal set of comparisons, for $K = 48$.

5.3 Hanna and CMCQRD

The previous experiments demonstrated that the PoE framework yields significant performance boosts in efficient settings. However, for the analyzed datasets, N is 16 and 6, and though PoE can

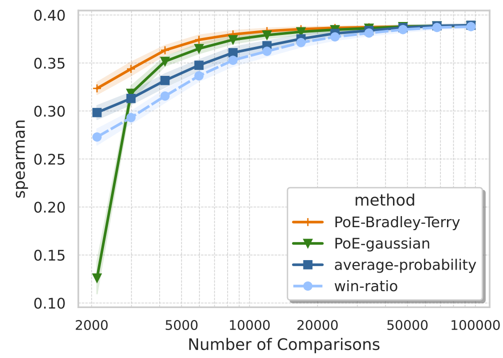


Figure 2: Mistral-7B, HANNA `COH`

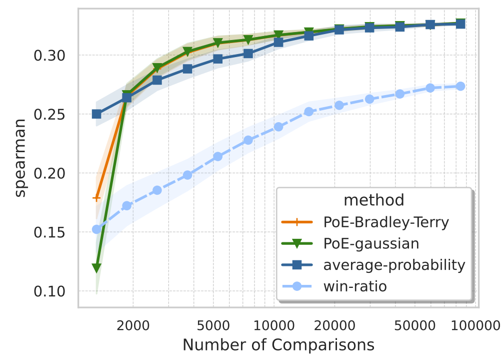


Figure 3: Llama2-13B, CMCQRD `DIF`

reduce the number of LLM calls, it is still feasible to run all $O(N^2)$ comparisons. This section now evaluates CMCQRD and HANNA, where N=1056 and N=658 respectively. Table 3 presents performance when using $\alpha \cdot N$ comparisons, where it's observed that POE-BT achieves consistently better performance than the average probability across all models and datasets. Faster convergence is observed for PoE-BT, with the average performance difference between 5 and 50 comparisons per item 0.8 SCC apart, while it is 2.5 SCC for the average probability. Note that evaluation was only conducted for Llama2 and Mistral due to FlanT5's maximum token length of 512.

Figure 3 illustrates the full efficiency curves for

system	K	CMCQRD DIF avg-prob	CMCQRD DIF PoE-BT	HANNA COH avg-prob	HANNA COH PoE-BT	HANNA CMP avg-prob	HANNA CMP PoE-BT	HANNA SUR avg-prob	HANNA SUR PoE-BT
Llama2-7B	$5N$	31.9	33.4	39.2	41.3	45.7	47.9	32.8	34.1
	$10N$	33.8	34.4	40.3	41.4	46.9	48.2	33.6	34.3
	$20N$	34.8	35.0	41.1	41.6	47.6	48.3	34.1	34.5
	$50N$	35.3	35.3	41.4	41.6	48.0	48.3	34.4	34.5
Llama2-13N	$5N$	30.0	31.2	39.9	41.3	51.7	54.6	34.6	36.9
	$10N$	31.5	31.9	41.2	41.8	53.4	54.9	36.0	37.2
	$20N$	32.2	32.3	41.8	41.9	54.3	55.1	36.8	37.5
	$50N$	32.6	32.6	42.1	42.1	54.9	55.1	37.2	37.6
Mistral-7B	$5N$	38.9	40.7	36.6	38.3	47.3	49.9	24.2	25.5
	$10N$	40.7	41.1	37.9	38.6	49.0	50.6	25.3	26.0
	$20N$	41.1	41.2	38.7	38.8	50.1	50.9	25.9	26.2
	$50N$	41.2	41.2	38.9	38.9	50.7	51.0	26.0	26.1

Table 3: Spearman correlations for CMCQRD and HANNA for specific attributes. $K \in \{5N, 10N, 20N, 50N\}$ is the total number of symmetric comparisons made, e.g., $5N$ refers to each sample being in 5 comparisons.

several models and attributes. We observe that PoE-BT typically performs best, and though PoE-g often performs similarly to PoE-BT, in very low information regions PoE-g can have poor correlations. In all cases, the PoE methods appear to mostly converge to their solution within $10 \cdot N$ comparisons, significantly fewer than $N(N-1)$.

5.4 Non-Symmetric Comparisons

Previously, to minimize the influence of positional bias and model inconsistency, both permutations of any comparison were evaluated. Although this reduces bias, one may gain more information by having a more diverse set of comparisons. Mistral-7B has minimal positional bias with $E[p_{ij}] = 0.51$, while Llama-7B has considerable bias with $E[p_{ij}] = 0.78$. To investigate whether symmetry is required, we look at performance of the non-symmetric set-up for Mistral-7B and Llama-7B (shown in Appendix Figure 7). For Llama2-7B, the debiased expert yields large performance gains while for Mistral-7B, the debiasing parameter has little influence, as expected since γ will be near 0. Note that, although Llama2-7B is more biased, it has better judgement capabilities and achieves better correlations, though the debiasing parameter is required. Figure 4 compares non-symmetric debiased performance with symmetric performance and illustrates that the two perform similarly, albeit with slightly different characteristics. Non-symmetric often does better in the low number of comparisons region, symmetric sometimes marginally better after, and performance is similar when more comparisons are made. Results for other models and attributes are presented in Appendix C.6.

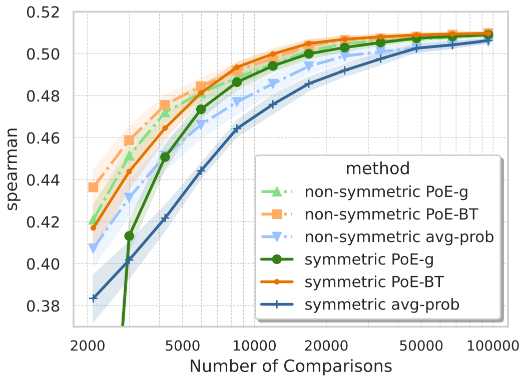


Figure 4: Mistral-7B, HANNA COH, symmetric vs non-symmetric

6 Conclusions

Comparative assessment using LLMs has been shown to be effective for text assessment. This paper investigates framing the scoring process within a Product of Experts framework, where the comparison information (including model confidence) can be easily combined to determine a set of scores that effectively capture text quality. This enables comparative assessment to not suffer from slow convergence rates, as now only a subset of the possible comparisons is used to predict the scores, but maintain the performance from when using the full set of comparisons. Further, using Gaussian experts yields a closed-form solution and provides a basis for deriving a greedy-optimal set of comparisons. The paper demonstrated the effectiveness of the approach on multiple different standard NLG evaluation datasets, such as SummEval and TopicalChat, as well as for large datasets where $N > 500$, which led to substantial computational savings against standard methods.

7 Limitations

The LLM comparisons can depend largely on the selected prompts used and the process used to extract probabilities. We chose simple prompts, but did not investigate the impact of prompt sensitivity and how well the approach holds when weaker/stronger prompts are used. Though due to the zero-shot nature, and the consistent observed performance boosts, our method to remain effective is likely in such settings, though this was not verified. Further, we are able to apply a soft-variant of Zermello to quickly optimise the PoE-Bradley-Taylor approach. However, when the bias term is introduced, soft-zero cannot be applied, and optimization of the solution is significantly slower. Nonetheless, since the main computational costs is associated with LLM calls, this is not a significant drawback. Lastly, our method is effective only when soft LLM probabilities are available, though for some APIs probabilities are not available and our method is less effective in bure binary decision cases.

8 Ethical Statement

Our paper addresses the cases of using more efficient use of LLMs when being used for NLG assessment. Although our work makes automatic assessment more practical and applicable to more settings, overly relying on automatic assessment may yield unintended consequences, especially when models have implicit biases that may discriminate against certain styles. Therefore as well as using automatic evaluation as useful metrics for text quality, it is useful to maintain human evaluation to ensure that systems to not unfairly penalize particular styles or properties which in general may be fine for the task.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

David Beaudoin and Tim Swartz. 2018. A computationally intensive ranking system for paired comparison data. *Operations Research Perspectives*, 5:105–112.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Manuela Cattelan. 2012. Models for paired comparison data: A review with emphasis on dependent data.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *Preprint*, arXiv:2402.10669.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. 2022. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805.

Cyril Chhun, Pierre Colombo, Fabian Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

László Csató. 2013. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research*, 21:783–803.

Herbert Aron David. 1963. *The method of paired comparisons*, volume 12. London.

Roger R Davidson and Peter H Farquhar. 1976. A bibliography on the method of paired comparisons. *Biometrics*, pages 241–252.

Otto Dykstra. 1956. A note on the rank analysis of incomplete block designs–applications beyond the scope of existing tables. *Biometrics*, 12(3):301–306.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long and Short Papers), pages 3806–3812.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19.

Geoffrey E. Hinton. 1999. Products of experts. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 1–6. IET.

David R Hunter. 2004. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *Preprint*, arXiv:2403.16950.

Adian Liusie, Yassir Fathullah, and Mark JF Gales. 2024a. Teacher-student training for debiasing: General permutation debiasing for large language models. *arXiv preprint arXiv:2403.13590*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024b. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Jordan J Louviere, David A Hensher, and Joffre D Swait. 2000. *Stated choice methods: analysis and applications*. Cambridge university press.

R Duncan Luce. 2005. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. *arXiv preprint arXiv:2301.12307*.

Charles F Manski. 1977. The structure of random utility models. *Theory and decision*, 8(3):229.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Tom Minka, Ryan Cleven, and Yordan Zaykov. 2018. Trueskill 2: An improved bayesian skill rating system. *Technical Report*.

Piotr Molenda, Adian Liusie, and Mark J. F. Gales. 2024. Waterjudge: Quality-detection trade-off when watermarking large language models. *Preprint*, arXiv:2403.19548.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark JF Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, et al. 2023. The cambridge multiple-choice questions reading dataset.

MEJ Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Vatsal Raina and Mark Gales. 2024. Question difficulty ranking for multiple-choice reading comprehension. *arXiv preprint arXiv:2404.10704*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

M. Welling. 2007. Product of experts. *Scholarpedia*, 2(10):3879. Revision #137078.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Heiga Zen, Mark JF Gales, Yoshihiko Nankaku, and Keiichi Tokuda. 2011. Product of experts for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):794–805.

Ernst Zermelo. 1929. Die berechnung der turnierergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Additional Theory for the Product of Expert Framework

A.1 Structure of $\tilde{\mathbf{W}}$ Matrix

The paper discussed the comparison matrix $\tilde{\mathbf{W}} \in R^{K+1 \times N}$, where each row represents the particular comparison being considered. It was discussed how for the k^{th} comparison between i and j, $\mathbf{W}_{ki} = 1$, $\mathbf{W}_{kj} = -1$, and $\mathbf{W}_{km} = 0 \;\; \forall m \neq i, j$. Further, an extra row was prepended to $\mathbf{W}$ adding constraints on the first score, forming $\tilde{\mathbf{W}}$ and ensuring the corresponding matrix is not defective. To illustrate the structure of $\tilde{\mathbf{W}}$, consider the case where one has 4 elements $x_{1:4}$ and all possible comparisons are considered,

$$\tilde{\mathbf{W}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{18}$$

A.2 Structure of $\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}$ Matrix

In the Gaussian-Products of Experts, the variance was shown to be directly related to the matrix $\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}$. For the full comparison case previously considered, this would yield a matrix of the form,

$$\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}} = \begin{bmatrix} 4 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} \tag{19}$$

Let $\tilde{\mathbf{A}} = \tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}$. For any set of selected comparisons, $\tilde{\mathbf{A}}_{ij} = \tilde{\mathbf{w}}_i \cdot \tilde{\mathbf{w}}_j$. Therefore by taking into account the structure of $\tilde{W}$, it's easily shown that the diagonal elements represent the number of comparisons the element has been involved in, while the off-diagonal elements are -1 if the comparison is made,

$$\tilde{\mathbf{A}}_{kk} = \sum_i 1(x_k \in \mathcal{C}_i) \tag{20}$$

$$\tilde{\mathbf{A}}_{ij} = \begin{cases} -1 & \text{if } (x_i, x_j) \in \mathcal{C}_K, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

This means that for the full comparison matrix, irrespective of N, the matrix $\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}$ will have the form,

$$\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}} = \begin{bmatrix} N & -1 & -1 & \dots & -1 \\ -1 & N-1 & -1 & \dots & -1 \\ -1 & -1 & N-1 & \dots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \dots & N-1 \end{bmatrix}$$

A.3 Equivalence of Gaussian PoE Solution with Average Probability

Given the structure of $\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}$, when considering the full-comparison set-up, the inverse is given by,

$$\left(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{-1} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1+\frac{2}{N} & 1+\frac{1}{N} & \dots & 1+\frac{1}{N} \\ 1 & 1+\frac{1}{N} & 1+\frac{2}{N} & \dots & 1+\frac{1}{N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1+\frac{1}{N} & 1+\frac{1}{N} & \dots & 1+\frac{2}{N} \end{bmatrix}$$

$$= \frac{N+1}{N} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix} + \frac{1}{2N} \begin{bmatrix} -1 & -1 & -1 & \dots & -1 \\ -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \dots & 1 \end{bmatrix}$$

For the Gaussian PoE with linear mean and constant Gaussian assumptions, the solution was shown to be of form $\hat{\mathbf{s}} = \alpha \cdot (\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}\tilde{\boldsymbol{\mu}}$. By noting that $\tilde{\boldsymbol{\mu}}$ represents the LLM probabilities for each comparative decision, we observe that $\tilde{\mathbf{W}}\tilde{\boldsymbol{\mu}}$ simply represents the sum of probabilities for all comparisons that each element has been a part of. Therefore, the above equation shows that the solution will be a constant shift of the average probability for any particular sample.

A.4 The Limiting Behaviour of the Bradley-Terry Model

Recall that the Bradley-Terry model, which uses discrete outcomes, has form

$$\mathtt{P}(\mathcal{C}_{1:K}|s_{1:N}) = \prod_{i,j \in \mathcal{C}_{1:K}} \mathtt{P}(y_{ij}|s_{1:N}) \tag{22}$$

$$\mathtt{P}(y_{ij}|s_{1:N}) = \sigma(s_i - s_j)^{y_{ij}}(1 - \sigma(s_i - s_j))^{1-y_{ij}}$$

Let us consider the situation where multiple outcomes of the same comparison are sampled from the LLM, assuming that each hard decision y_{ij} is drawn from Bernoulli distribution such that $y_{ij} \sim \text{Bernoulli}(p_{ij})$. One can define $C_{1:K}^{(i,j)}$ as all the comparisons sampled between x_i and x_j. The log probability of the comparisons can then be decomposed as,

$$
\begin{aligned}
&\log \mathtt{P}(\mathcal{C}_{1:K}|s_{1:N}) && (23)\\
=&\sum_{i,j,y_{ij}} \log \mathtt{P}(y_{ij}|s_{1:N}) && (24)\\
=&\sum_{i}\sum_{j}\sum_{y_{ij}\in C_{1:K}^{(i,j)}} \log \mathtt{P}(y_{ij}|s_{1:N}) && (25)\\
=&\sum_{i}\sum_{j} M \cdot \frac{1}{M}\sum_{y_{ij}\in C_{1:K}^{(i,j)}} \log \mathtt{P}(y_{ij}|s_{1:N}) && (26)
\end{aligned}
$$

Where $M \in \mathbb{R}$. However, let M represent the number of times each comparisons is made, such that $|C_{1:K}^{(i,j)}| = M$. By considering the limiting case where $M \to \infty$, the expression will then tend to,

$$
\begin{aligned}
&\frac{1}{M}\sum_{y_{ij}\in C_{1:K}^{(i,j)}} \log \mathtt{P}(y_{ij}|s_{1:N})\\
=&\frac{1}{M}\sum_{y_{ij}\in C_{1:K}^{(i,j)}} y_{ij}\log\sigma(s_i\text{-}s_j)+(1\text{-}y_{ij})\log(1\text{-}\sigma(s_i\text{-}s_j))\\
=&\mathbb{E}_{y_{ij}}[y_{ij}\log\sigma(s_i-s_j)+(1-y_{ij})\log(1-\sigma(s_i-s_j))]\\
=&p_{ij}\log\sigma(s_i-s_j)+(1-p_{ij})\log(1-\sigma(s_i-s_j))
\end{aligned}
$$

Therefore as $M \to \infty$,

$$
\begin{aligned}
&\sqrt[M]{\mathtt{P}(\mathcal{C}_{1:K}|s_{1:N})} && (27)\\
=&\prod_{i,j,p_{ij}\in\mathcal{C}_{1:K}} \sigma(s_i-s_j)^{p_{ij}}(1-\sigma(s_i-s_j))^{1-p_{ij}} && (28)
\end{aligned}
$$

A.5 Form of the Gaussiam PoE Score Distribution

Given $\mathtt{p}(\mathbf{W}\mathbf{s}|\mathcal{C}_{1:K}) = \mathcal{N}\Big(\mathbf{W}\mathbf{s};\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}\Big)$, to determine $\mathtt{p}(\mathbf{s}|\mathcal{C}_{1:K})$ one can expand the expression and isolate all terms that have an $\mathbf{s}$, yielding,

$$
\begin{aligned}
&\mathtt{p}(\mathbf{W}\mathbf{s}|\mathcal{C}_{1:K}) && (29)\\
=&\mathcal{N}\Big(\mathbf{W}\mathbf{s};\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}\Big) && (30)\\
\propto&\exp\left(-\frac{1}{2}\,(\mathbf{W}\mathbf{s}-\tilde{\boldsymbol{\mu}})^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{W}\mathbf{s}-\tilde{\boldsymbol{\mu}})\right) && (31)\\
\propto&\exp\left(-\frac{1}{2}\left(\mathbf{s}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{W}\mathbf{s}+2\mathbf{s}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right)\right)
\end{aligned}
$$

As the distribution over scores will be Gaussian, $\mathtt{p}(\mathbf{s}|C_{1:K}) \sim \mathcal{N}(\mathbf{s};\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$, one can equate coefficients to derive the form used in the paper,

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_s^* &= (\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})^{-1} && (32)\\
\boldsymbol{\mu}_s^* &= (\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} && (33)
\end{aligned}
$$

Which has pdf,

$$
\frac{1}{(2\pi)^{N/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{s}-\boldsymbol{\mu}_s^*)^{\mathsf{T}}\boldsymbol{\Sigma}^{*-1}(\mathbf{s}-\boldsymbol{\mu}_s^*)\right)
$$

The maximum probability scores will be at the mean, $\mathbf{s} = \boldsymbol{\mu}_s^*$, which has a probability of,

$$
\begin{aligned}
&\frac{1}{(2\pi)^{N/2}\mathtt{det}\left((\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})^{-1}\right)^{1/2}} && (34)\\
=&\frac{\sqrt{\mathtt{det}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{W}})}}{(2\pi)^{N/2}} && (35)
\end{aligned}
$$

For the linear Gaussian, where it is assumed that $\tilde{\boldsymbol{\Sigma}} = \sigma^2\mathbf{I}$, this can be reduced to,

$$
\mathtt{p}(\mathbf{s}{=}\boldsymbol{\mu}_s^*|\mathcal{C}_{1:K}) = \frac{\sqrt{\mathtt{det}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}})}}{(2\pi\sigma^2)^{N/2}} \quad (36)
$$

A.6 Efficient Greedy Comparison Selection

Assume that $\tilde{\mathbf{W}}^{(k)*}$ is the selected comparison matrix using k comparisons. Considering an additional comparison (i,j) is equivalent to adding an extra row $\mathbf{r} \in R^N$ where $\mathbf{r}_i = 1$, $\mathbf{r}_j = -1$ and $\mathbf{r}_l{=}0\ \ \forall l \neq i,j$. By noting that,

$$
\begin{aligned}
&\mathtt{det}\left([\tilde{\mathbf{W}};\mathbf{r}]^{\mathsf{T}}[\tilde{\mathbf{W}};\mathbf{r}]\right) && (37)\\
=&\mathtt{det}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}+\mathbf{r}\mathbf{r}^{\mathsf{T}}) && (38)\\
=&\mathtt{det}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}})(1+\mathbf{r}^{\mathsf{T}}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}})^{-1}\mathbf{r}) && (39)
\end{aligned}
$$

the next optimal comparison $(\hat{i},\hat{j})$ is calculated as,

$$
\hat{i},\hat{j} = \arg\max_{i,j} \mathbf{A}_{ii}^{(k)*} + \mathbf{A}_{jj}^{(k)*} - 2\cdot\mathbf{A}_{ij}^{(k)*} \quad (40)
$$

Updating $\tilde{\mathbf{W}}^{(k)*}$ is trivial, since considering an additional comparison (i,j) is equivalent to adding an extra row $\mathbf{r} \in R^N$ to $\tilde{\mathbf{W}}^{(k)*}$, where $\mathbf{r}_i = 1$, $\mathbf{r}_j{=}-1$ and $\mathbf{r}_l{=}0\ \ \forall l \neq i,j$. Therefore

$$
\tilde{\mathbf{W}}^{(k+1)*} = [\tilde{\mathbf{W}}^{(k)*};\mathbf{r}] \quad (41)
$$

However one can also efficiently update the inverse using the Sherman-Morrison inversion lemma,

$$\mathbf{A}^{(k+1)*} = \left([\tilde{\mathbf{W}}^{(k)*}; \mathbf{r}]^\mathsf{T} [\tilde{\mathbf{W}}^{(k)*}; \mathbf{r}]\right)^{-1} \tag{42}$$

$$= \left(\tilde{\mathbf{W}}^{(k)*\mathsf{T}} \tilde{\mathbf{W}}^{(k)*} + \mathbf{r}\mathbf{r}^\mathsf{T}\right)^{-1} \tag{43}$$

$$= \mathbf{A}^{(k)*} - \frac{\mathbf{A}^{(k)*}\mathbf{r}\mathbf{r}^\mathsf{T}\mathbf{A}^{(k)*}}{1 + \mathbf{r}^\mathsf{T}\mathbf{A}^{(k)*}\mathbf{r}} \tag{44}$$

Note that to initialize $\tilde{\mathbf{W}}$, the simplest option would be to use $N-1$ comparisons and follow a stripped diagonal matrix, e.g.

$$\tilde{\mathbf{W}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{45}$$

A.7 Detailed Derivation of β for the Debiased PoE-Gaussian Expert

For a given expert, $\mathrm{p}(s_i - s_j|p_{ij})$, and an underlying LLM which generates comparative decisions, $\mathrm{p}_{\text{LLM}}(p_{ij})$ (assuming the underlying texts x_i and x_j are randomly drawn), there is an associated marginalised distribution of score differences, $\mathrm{p}(s_i - s_j)$. Note that as the texts are randomly drawn, they are equally likely to be drawn in either position and therefore, $\mathbb{E}[s_i - s_j] = 0$. For a debiased expert $\mathrm{p}_\gamma(s_i - s_j|p_i j)$, the objective is to find the parameter γ for the LLM that ensures that $\mathbb{E}[s_i - s_j] = 0$,

$$\mathbb{E}[s_i - s_j] \tag{46}$$

$$= \int_{-\infty}^{\infty} (s_i\text{-}s_j)\mathrm{p}(s_i\text{-}s_j)d(s_i\text{-}s_j) \tag{47}$$

$$= \int_0^1 \int_{-\infty}^{\infty} (s_i\text{-}s_j)\mathrm{p}_\gamma(s_i\text{-}s_j|p_{ij})\mathrm{p}_{\text{LLM}}(p_{ij})d(s_i\text{-}s_j)dp_{ij}$$

$$= \int_0^1 \mathrm{p}_{\text{LLM}}(p_{ij}) \int_{-\infty}^{\infty} (s_i\text{-}s_j)\mathrm{p}_\gamma(s_i\text{-}s_j|p_{ij})d(s_i\text{-}s_j)dp_{ij}$$

$$= \int_0^1 \mathrm{p}_{\text{LLM}}(p_{ij}) \cdot \mathbb{E}[s_i - s_j|p_{ij}, \gamma]\ dp_{ij} \tag{48}$$

The parameter γ was proposed to be a simple linear shift of the score differences, such that $\mathrm{p}_\gamma(s_i\text{-}s_j|p_{ij}) = \mathrm{p}(s_i\text{-}s_j\text{-}\gamma|p_{ij})$. For the linear Gaussian, $\mathcal{N}\big(s_i\text{-}s_j; \alpha\cdot(p_{ij}\text{-}\beta), \sigma^2\big)$ this is equivalent to setting the β parameter. The mean of the expert is $\alpha\cdot(p_{ij}\text{-}\beta)$, and therefore,

$$\mathbb{E}[s_i\text{-}s_j] = \int_0^1 \mathrm{p}_{\text{LLM}}(p_{ij}) \cdot \mathbb{E}[s_i\text{-}s_j|p_{ij}]\ dp_{ij} \tag{49}$$

$$= \int_0^1 \mathrm{p}_{\text{LLM}}(p_{ij}) \cdot \alpha\cdot(p_{ij}\text{-}\beta)\ dp_{ij} \tag{50}$$

$$= \alpha\left(\int_0^1 p_{ij}\ \mathrm{p}_{\text{LLM}}(p_{ij})\ dp_{ij} - \beta\right) \tag{51}$$

Which setting to zero yields $\beta = \mathbb{E}[p_{ij}] \approx \frac{1}{K}\sum_{k=1}^{K} p_{ij}^{(k)}$, i.e. β should be set to the average LLM probability.

A.8 Deriving γ for the Debiased PoE-BT Expert

For experts that are unstable or for which the expectation is analytically intractable, one can instead ensure the mode of the skill difference likelihood is set to 0 when the skill difference is 0. Differentiating the expected score difference yields,

$$\frac{\partial}{\partial\gamma}\mathbb{E}[\log \mathrm{p}_\gamma(s_i - s_j)] \tag{52}$$

$$= \frac{\partial}{\partial\gamma}\int_0^1 \log \mathrm{p}_\gamma(s_i\text{-}s_j|p_{ij})\mathrm{p}(p_{ij})dp_{ij} \tag{53}$$

$$= \int_0^1 \mathrm{p}_{\text{LLM}}(p_{ij})\frac{\partial}{\partial\gamma}\Big(\log \mathrm{p}_\gamma(s_i\text{-}s_j|p_{ij})\Big)dp_{ij} \tag{54}$$

The probabilistic Bradley-Terry accounting for bias has form,

$$\mathrm{p}_\gamma(s_i - s_j|p_{ij}) = \frac{1}{Z_{ij}} \cdot \frac{e^{p_{ij}\cdot(s_i - s_j - \gamma)}}{1 + e^{(s_i - s_j - \gamma)}} \tag{55}$$

which when differentiated yields,

$$\frac{\partial}{\partial\gamma}\log \mathrm{p}(s_i - s_j|p) \tag{56}$$

$$= \frac{\partial}{\partial\gamma}\big(p_{ij}\cdot(s_i\text{-}s_j\text{-}\gamma) - \log(1 + e^{s_i - s_j - \gamma})\big) \tag{57}$$

$$= -\,p_{ij} + \frac{e^{s_i - s_j - \gamma}}{1 + e^{s_i - s_j - \gamma}} \tag{58}$$

Evaluating the integral at $s_i - s_j = 0$,

$$\frac{\partial}{\partial\gamma}\mathbb{E}[\log \mathrm{p}_\gamma(s_i - s_j)]\Big|_{s_i - s_j = 0} \tag{59}$$

$$= \int_0^1 \mathrm{p}_{\text{LLM}}(p_{ij})\left(p_{ij} + \frac{e^{-\gamma}}{1 + e^{-\gamma}}\right)dp_{ij} \tag{60}$$

setting to zero yields, $\gamma = -1\cdot\log\left(\frac{\mathbb{E}[p_{ij}]}{1+\mathbb{E}[p_{ij}]}\right) = -\text{logit}(\mathbb{E}[p_{ij}]) \approx \text{logit}\left(\frac{1}{K}\sum_{k=1}^{K} p_{ij}^{(k)}\right)$

dataset	score	prompt
SummEval	`COH`	Article: <context>\n\nSummary A: <A> \n\nSummary B: <B> \n\nWhich Summary is more coherent, Summary A or Summary B?
SummEval	`CON`	Article: <context> \n\nSummary A: <A> \n\nSummary B: <B> \n\nWhich Summary is more consistent to the article, Summary A or Summary B?
TopicalChat	`CNT`	Dialogue: <context> \n\nResponse A: <A> \n\nResponse B: <B> \n\nWhich Response continues the dialogue better, Response A or Response B?
TopicalChat	`NAT`	Dialogue: <context> \n\nResponse A: <A> \n\nResponse B: <B> \n\nWhich Response appears more natural, Response A or Response B?
HANNA	`SUR`	Story A: \n<A> \n\nStory B: \n<B> \n\nWhich story is more surprising, Story A or Story B?
HANNA	`CMP`	Story A: \n<A> \n\nStory B: \n<B> \n\nWhich story is more complex, Story A or Story B?
CMCQRD	`DIF`	Question A: \n<A> \n\nQuestion B: \n<B> \n\nWhich reading comprehension question is more difficult to answer, Question A or Question B?

Table 4: Prompts used for prompting the LLM to make pairwise decisions between two candidate texts.

B Experimental Details

B.1 Prompts

Table 4 shows examples of the prompts used for generating comparative decisions (other prompts for other attributes were of similar style). For a particular dataset and attribute, all models are provided with the same simple prompts, which were the only prompts used for experiments. No prompt engineering was done, matching situations where one doesn't have access to labels to evaluate systems.

B.2 Computation Resources

All experiments were run on L40 machines, where evaluation was parallelised over 4 machines. Each SummEval attribute took a 1 L40 GPU hours for Llama2-7b, Mistral-7B, and FlanT5-3B (despite being smaller, FlanT5 is float32 and hence not faster) while Llama2-13B took 2 hours and FlanT5-11B took 2.5 hours. For each attribute of HANNA, performing 200,000 comparisons required 8/8/9/15/21 GPU hours for Llama2-7B/Mistral-7B/FlanT5-3B/Llama2-13B/FlanT5-11B. For CMCQRD performing 200,000 comparisons required 8/8/9/15/21 GPU hours for Llama2-7B/Mistral-7B/FlanT5-3B/Llama2-13B/FlanT5-11B. All TopicalChat experiments could be run in under 30 minutes.

B.3 Model and Dataset Licences

Model Licenses: LLaMA-2-7B-chat and LLaMA-2-13B-chat (Touvron et al., 2023) use a LLaMA-2 license. Mistral-7B-Instruct-v0.2 uses an Apache-2.0 license. Similarly, FlanT5-3B and FlanT5-11B use an Apache-2.0 license.

Dataset Licenses: SummEval (Fabbri et al., 2021) uses an MIT License. TopicalChat (Mehri and Eskenazi, 2020) uses the MIT License. Hanna (Chhun et al., 2022) uses an MIT License. CMCQRD (Mullooly et al., 2023) uses its own license.

C Additional Results

C.1 SummEval Pearson Performance Tables

The main paper illustrated the context-level Spearman correlations for SummEval, which Table 5 also shows the standard deviations of. For certain applications, one may not only care about the rank ordering of the points but also the relative spacing between them, as this provides information on the predicted quality difference between any two texts. Table 6 therefore presents the Pearson correlations for SummEval, where similar trends to the Spearman table are observed.

		decisions only			probabilities		
system	K	win-ratio	BT	PoE-g-hard	avg-prob	PoE-BT	PoE-g
Llama2-7B	48	21.6±0.8	23.4±0.7	22.5±0.7	24.0±0.7	26.8±0.5	26.6±0.5
	240	27.8±0.0	27.9±0.0	27.6±0.0	28.4±0.0	28.4±0.0	28.4±0.0
Llama2-13B	48	30.8±0.7	33.1±0.7	31.6±0.7	33.7±0.6	37.7±0.4	37.3±0.4
	240	39.3±0.0	39.3±0.0	39.2±0.0	39.3±0.0	39.3±0.0	39.3±0.0
Mistral-7B	48	29.7±0.8	31.9±0.7	30.5±0.6	31.1±0.7	33.2±0.6	32.8±0.6
	240	38.1±0.0	38.1±0.0	38.0±0.0	37.7±0.0	37.7±0.0	37.7±0.0
FlanT5-3B	48	34.1±0.8	36.6±0.6	34.9±0.7	38.4±0.6	42.6±0.4	42.4±0.4
	240	43.6±0.0	43.6±0.0	43.4±0.0	44.3±0.0	44.3±0.0	44.3±0.0
FlanT5-11B	48	31.2±0.8	33.4±0.7	32.0±0.7	34.7±0.7	38.5±0.4	38.4±0.4
	240	40.0±0.0	40.0±0.0	39.7±0.0	40.5±0.0	40.5±0.0	40.5±0.0

Table 5: Spearman Correlations for SummEval, averaged over all attributes (COH, CON, FLU, REL). K is the number of comparisons made, where $K = 240$ is the full set of comparisons.

system	R	win-ratio	BT	PoE-g-hard	avg-prob	PoE-BT	PoE-g
Llama2-7B	48	21.7±0.7	23.5±0.6	22.3±0.7	24.3±0.6	26.9±0.5	26.8±0.4
	240	27.8±0.0	27.8±0.0	27.8±0.0	28.4±0.0	28.4±0.0	28.4±0.0
Llama2-13B	48	31.3±0.7	33.8±0.6	32.0±0.7	36.0±0.5	40.6±0.3	39.9±0.4
	240	39.8±0.0	40.4±0.0	39.9±0.0	42.1±0.0	42.5±0.0	42.1±0.0
Mistral-7B	48	30.8±0.7	33.3±0.7	31.6±0.6	32.5±0.6	35.5±0.7	34.7±0.7
	240	39.7±0.0	40.5±0.0	39.7±0.0	39.9±0.0	41.3±0.0	39.9±0.0
FlanT5-3B	48	34.3±0.8	37.2±0.7	35.0±0.7	42.3±0.5	48.3±0.3	47.1±0.3
	240	44.1±0.0	45.0±0.0	44.1±0.0	49.4±0.0	50.0±0.0	49.4±0.0
FlanT5-11B	48	31.7±0.7	34.2±0.7	32.3±0.7	37.3±0.6	41.8±0.5	41.4±0.5
	240	40.8±0.0	41.4±0.0	40.8±0.0	43.7±0.0	44.0±0.0	43.7±0.0

Table 6: Pearson correlations for SummEval, averaged over all attributes (COH, CON, FLU, REL). K is the number of balanced comparisons made, where $K = 120$ is the full set of comparisons.

C.2 TopicalChat Performance Tables

Table 7 and 8 demonstrate performance for comparative assessment when applied to dialogue evaluation. The PoE approaches continue to provide considerable performance improvements at the operating point $K=18$, albeit since N is not very large ($N=6$), the full set of comparisons is only 30 comparisons and fairly feasible to compute, and so for these experiments the computational savings are less significant.

system	R	win-ratio	BT	PoE-g-hard	avg-prob	PoE-BT	PoE-g
Llama2-7B	18	28.4±1.2	28.9±1.0	28.7±1.1	27.7±1.4	29.7±0.9	29.5±1.0
	30	31.5±0.0	31.6±0.0	31.6±0.0	31.5±0.0	31.5±0.0	31.5±0.0
Llama2-13B	18	37.4±1.1	38.1±1.1	37.9±1.0	38.4±1.2	40.5±0.8	40.5±0.9
	30	41.6±0.0	41.7±0.0	41.8±0.0	41.6±0.0	41.6±0.0	41.6±0.0
Mistral-7B	18	42.8±1.1	43.3±0.9	43.2±1.3	42.8±1.2	45.3±1.1	44.8±1.0
	30	47.4±0.0	47.2±0.0	47.7±0.0	46.9±0.0	46.9±0.0	46.9±0.0
FlanT5-3B	18	41.3±1.3	41.8±1.2	41.6±1.3	43.4±1.2	45.4±0.8	45.2±0.8
	30	45.3±0.0	44.8±0.0	45.3±0.0	44.7±0.0	44.7±0.0	44.7±0.0
FlanT5-11B	18	51.2±1.2	52.4±1.1	51.9±1.1	53.8±1.1	56.2±0.8	56.1±0.8
	30	57.0±0.0	56.6±0.0	56.0±0.0	58.1±0.0	58.1±0.0	58.1±0.0

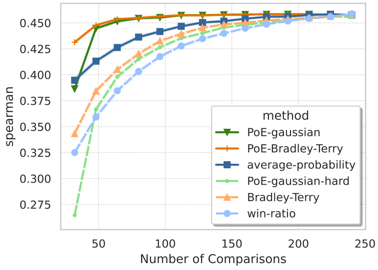
Table 7: Spearman correlations for TopicalChat, averaged over all attributes (COH, CNT, ENG, NAT). K is the number of comparisons made, where $K=30$ is the full set of comparisons.

system	R	win-ratio	BT	PoE-g-hard	avg-prob	PoE-BT	PoE-g
Llama2-7B	18	28.5±1.1	29.4±0.8	29.1±1.0	29.1±1.1	29.4±0.8	30.2±0.7
	30	31.6±0.0	31.6±0.0	31.6±0.0	31.5±0.0	30.7±0.0	31.5±0.0
Llama2-13B	18	37.5±1.1	38.7±1.0	38.4±1.0	40.2±1.0	41.8±0.5	41.8±0.6
	30	41.4±0.0	41.5±0.0	41.4±0.0	42.5±0.0	42.6±0.0	42.5±0.0
Mistral-7B	18	42.0±1.1	43.2±0.9	43.0±1.2	44.4±1.0	46.1±0.9	46.1±0.7
	30	46.4±0.0	46.3±0.0	46.4±0.0	48.1±0.0	48.4±0.0	48.1±0.0
FlanT5-3B	18	42.1±1.2	43.1±1.1	42.8±1.1	45.7±1.0	48.0±0.7	47.9±0.7
	30	46.5±0.0	46.5±0.0	46.5±0.0	48.7±0.0	48.6±0.0	48.7±0.0
FlanT5-11B	18	51.5±1.2	53.3±1.0	52.9±1.0	56.3±0.9	58.1±0.6	58.3±0.6
	30	57.5±0.0	57.4±0.0	57.4±0.0	59.8±0.0	59.7±0.0	59.8±0.0

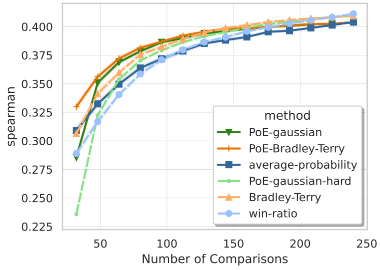
Table 8: Pearson correlations for TopicalChat averaged over all attributes (COH, CNT, ENG, NAT). K is the number of comparisons made, where $K=30$ is the full set of comparisons.

C.3 SummEval and Topical Chat Efficiency Plots

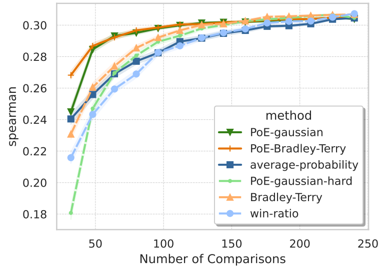
Figure 5 showcases the performance of the various scoring approaches for further models/attributes for SummEval and TopicalChat. We observe that in all cases the PoE approaches lead to best performance when only a subset of comparisons are used.



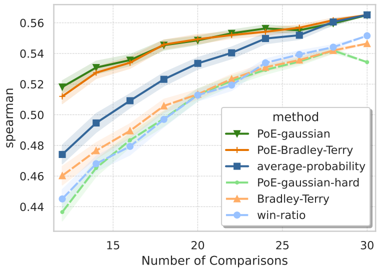
(a) FlanT5-3B, SummEval `CON`



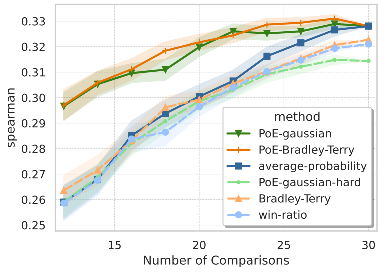
(b) Mistral-7B, SummEval `COH`



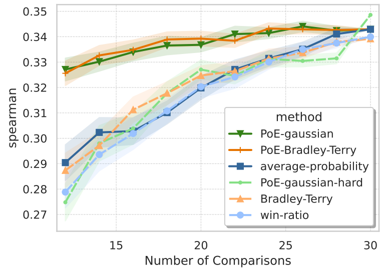
(c) Llama-13B, SummEval `FLU`



(d) FlanT5-11B, TopicalChat `COH`



(e) Llama-7B, TopicalChat `CNT`

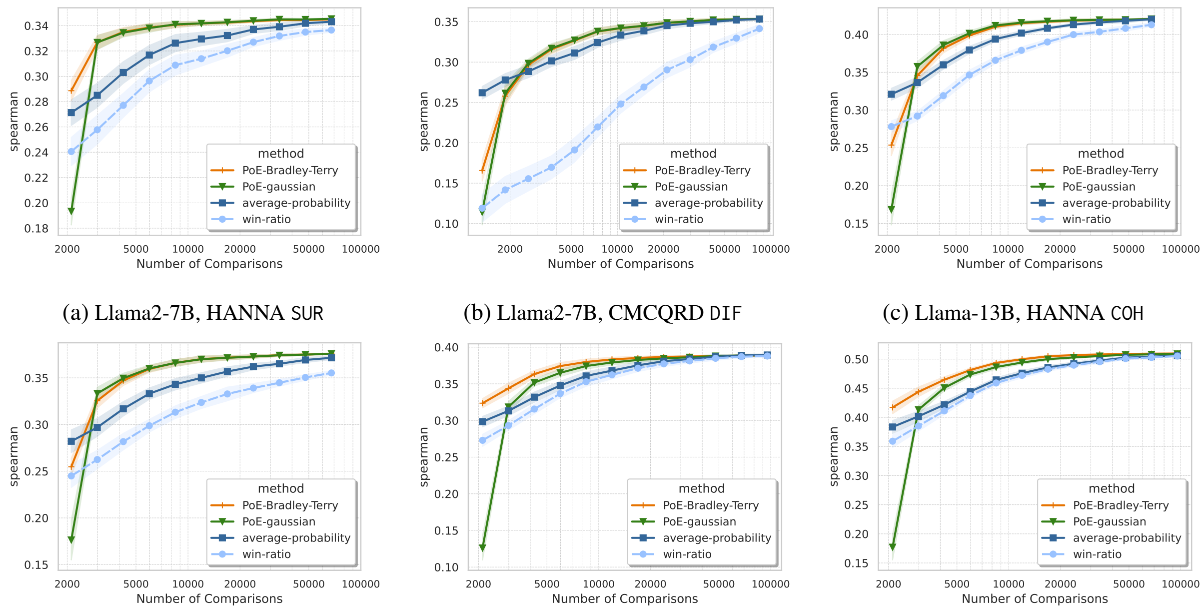


(f) Llama2-13B, TopicalChat `NAT`

Figure 5: Efficiency curves when sweeping K, the number of comparisons per context, where at each K the comparisons are randomly drawn 100 times. Average performance with 95% confidence is displayed. These curves were randomly selected from all possible configurations.

C.4 HANNA and CMCQRD Chat Efficiency Plots

Figure 6 showcases further performance curves for HANNA and CMCQRD, which demonstrate the effectiveness of the PoE framework in further settings with large N.

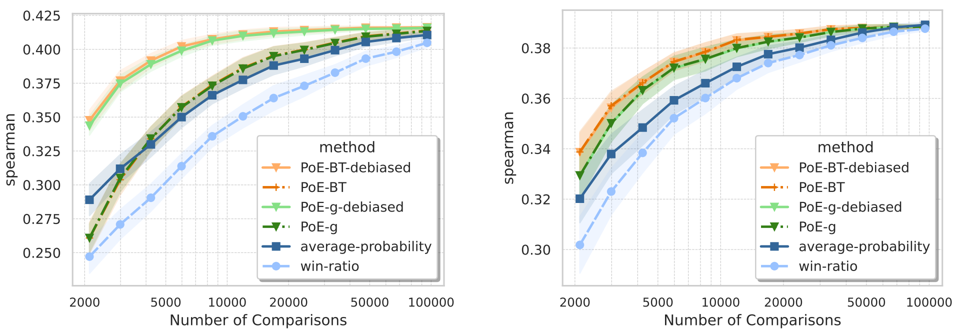


(a) Llama2-7B, HANNA SUR
(b) Llama2-7B, CMCQRD DIF
(c) Llama-13B, HANNA COH
(d) Llama-13B, HANNA SUR
(e) Mistral-7B, HANNA COH
(f) Mistral-7B, HANNA CMP

Figure 6: Efficiency curves where comparisons are randomly drawn 20 times. These curves were randomly selected from all possible configurations.

C.5 Non-Symmetric Efficiency Plots

Figure 7 shows the performance curves for Llama-7B and Mistral 7B. Mistral-7B has minimal positional bias with $E[p_{ij}]=0.51$, while Llama-7B has considerable bias with $E[p_{ij}]=0.78$. For Llama2-7B, the debiased experts, $\mathrm{p}_{\gamma}(s_i - s_j|p_{ij})$, yield large performance gains and performance does not converge quickly without it. For Mistral-7B, the debiasing parameter has little influence, as expected since γ will be near 0. Note that, although Llama2-7B is more biased, it has better judgement capabilities and achieves better correlations, though the debiasing parameter is required.

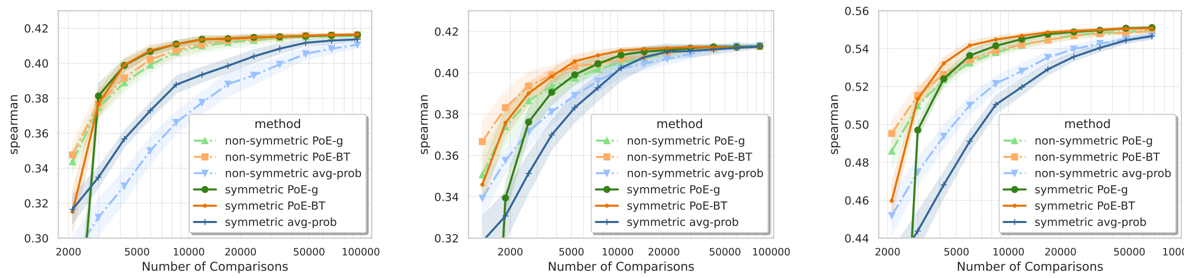


(a) Llama-7B, HANNA COH, non-symmetric
(b) Mistral-7B, HANNA COH, symmetric vs non-symmetric

Figure 7: Efficiency curves in the non-symmetric set-up.

C.6 Symmetric vs Non-Symmetric Efficiency Plots

For several other models and datasets, Figure 8 compares the performance between symmetric and non-symmetric attributes, as well as against the average probability and win-ratio. We observe that both perform well and often similarly, although minor differences in characteristics can be observed, as discussed in the main paper.



(a) Llama2-7B, HANNA COH, symmetric vs non-symmetric

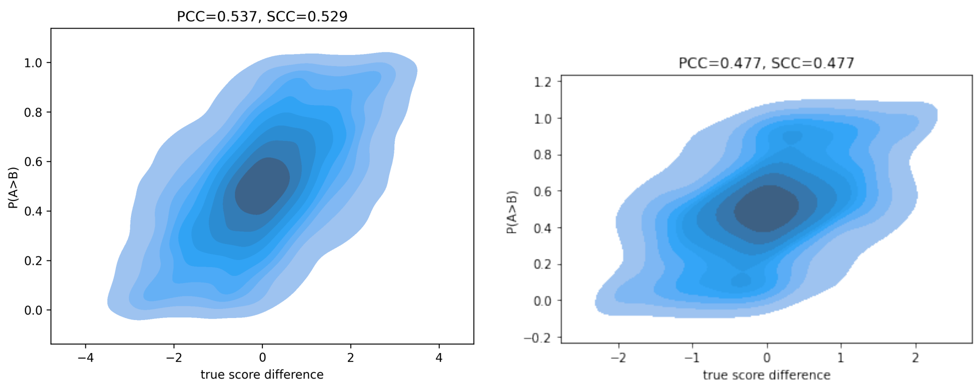
(b) Mistral-7B, CMCQRD DIF, symmetric vs non-symmetric

(c) Llama2-13B, HANNA CMP, symmetric vs non-symmetric

Figure 8: Efficiency Curves when sweeping K, the number of comparisons per context, with 95% confidence intervals using 100 samples per step for non-symmetric set-up. These curves were randomly selected from all possible configurations.

C.7 Data Analysis

In the POE framework, each expert models the distribution $\mathrm{p}(s_i - s_j | p_{ij})$. To determine a suitable form of the expert, and whether the Gaussian and/or the extended Bradley-Terry experts are sensible assumptions, Figure 9 displays the joint bivariate distribution between the true score difference $s_i - s_j$ and the observed probability p_{ij}. For a particular LLM, all comparisons over all the contexts of the dataset are assessed. The frequency count of the LLM probability and true score difference (calculated using the gold-standard annotator labels) is then plotted. The plots illustrate a clear correlation between the probabilities and score difference, implying that considerable scoring information can be gained from leveraging probabilities and decisions. However, the mapping is not deterministic, and there is considerable noise present. Empirically, The distributions appear to be well approximated by Gaussian distributions, implying that the conditional distributions will also be well-modelled by Gaussian distributions.

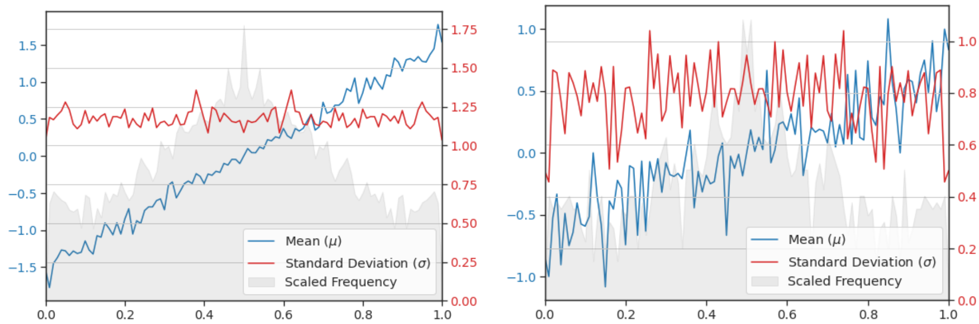


(a) FlanT5-3B, SummEval COH

(b) Llama2-13B, TopicalChat CNT

Figure 9: Joint distribution of the LLM probabilities and true scores.

We further analyze the relationship between the LLM probability p and the expected score difference, $\delta(p) = E_{p_{ij}}[s_i - s_j \mid |p_{ij} - p| < \epsilon]$. Figure 10 demonstrates that 1) the probability is quite linearly correlated with the expected score difference; and 2) the variance across all score distributions given the probability is quite constant. Therefore the Gaussian assumptions discussed in Section 3.4 appear to be reasonable.



(a) FlanT5-3B, SummEval `COH` (b) Llama2-13B, TopicalChat `CNT`

Figure 10: Expected score difference and variance given the LLM probability.

Note that TopicalChat is a smaller dataset (with 1800 total comparisons) and hence has more observed noise.

C.8 Comparison Against Additional baselines

Throughout the paper, baselines such as the Bradley Terry, average probability and win-ratio were used as methods to compare the best method to get scores from comparative outcomes. However alternate methods are possible, which do not necessarily combine information from a subset of the comparisons. For example, G-EVAL (Liu et al., 2023b) uses a prompt that asks the model to directly score texts and then calculates the fair mean over the probabilities of scores. While PairS (Liu et al., 2024) considers sorting algorithms to guide which pairwise comparisons should be made, as well as for determining the final rankings. Table 9 displays the performance of our Product of Experts Framework of LLM comparative assessment against these baselines for SummEval and HANNA (using a modest $K = 3N$ and $K = 5N$ respectively) and demonstrates that our approach has considerably better performance over the other baseline methods, where in 11/14 settings has the best performance (and often by considerable margins).

		SummEval				HANNA		
	K	COH	CON	FLU	REL	COH	CMP	SUR
Llama2-7B	G-Eval	15	23	7	20	25	33	17
	PAIRS-beam	17	**31**	18	24	29	17	19
	PoE-BT	**29**	24	**20**	**34**	**41**	**48**	**34**
Mistral-7B	G-Eval	25	**39**	20	25	34	39	25
	PAIRS-beam	28	30	24	27	33	31	**27**
	PoE-BT	**34**	36	**26**	**37**	**38**	**50**	26

Table 9: SummEval performance for SummEval and HANNA for all particular attributes.